feature pyramid networks for object detection

Tsung-Yi Lin et.al 2017

wyuzyf March 2019

1 论文概述

作者提出了多尺度 object detection 算法：FPN(feature pyramid network)。原来的 object detection 算法都只是采用顶层的特征做预测，但是我们知道低层的特征语义信息比较少，但是目标位置准确；高层的特征语义信息丰富，但是目标位置比较粗略. 也有些算法采用多尺度特征融合的方式，但是一般采用融合后的特征做预测。而本文不一样的地方在于预测是在不同特征层独立进行的。

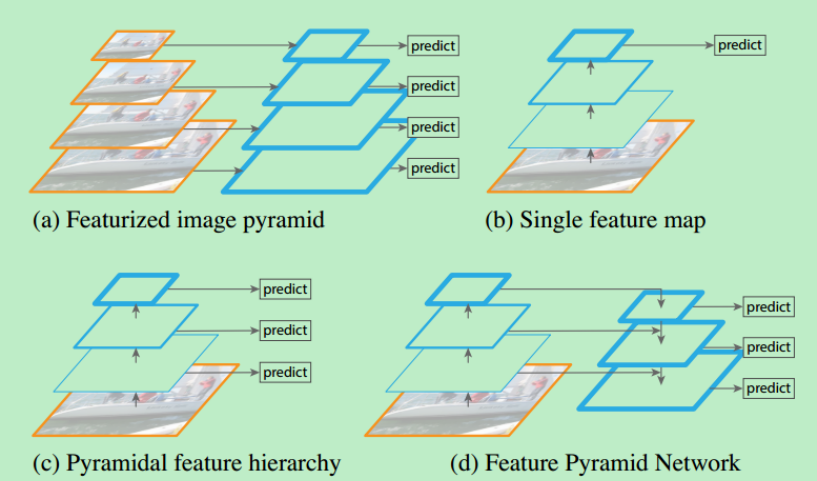

Figure 1. (a) Using an image pyramid to build a feature pyramid. Features are computed on each of the image scales independently, which is slow. (b) Recent detection systems have opted to use only single scale features for faster detection. (c) An alternative is to reuse the pyramidal feature hierarchy computed by a ConvNet as if it were a featurized image pyramid. (d) Our proposed Feature Pyramid Network (FPN) is fast like (b) and (c), but more accurate. In this figure, feature maps are indicate by blue outlines and thicker outlines denote semantically stronger features.

图 1: 利用特征的 4 种形式

2 论文详解

上图 1 展示了 4 种特征的形式：

(a) 图像金字塔，即对图片做 scale 操作，然后不同的 scale 图片做特征提取，对每级的特征做 predict。这种方式计算量相对较大；

(b) 像 SPP net，Fast RCNN，Faster RCNN 是采用这种方式，即仅采用网络最后一层的特征；

(c) 像 SSD(Single Shot Detector) 采用这种多尺度特征融合的方式，没有上采样过程，即从网络不同层抽取不同尺度的特征做预测，这种方式不会增加额外的计算量。作者认为 SSD 算法中没有用

到足够低层的特征（在 SSD 中，最低层的特征是 VGG 网络的 conv4-3), 而在作者看来足够低层的特征对于检测小物体是很有帮助的。

(d) 本文作者采用的形式。同 (c) 中的方法有些类似，也是拿单一维度的图片作为输入，然后它会选取所有层的特征来处理然后再联合起来做为最终的特征输出组合。（作者在论文中拿 Resnet 为实例时并没选用 Conv1 层，那是为了算力及内存上的考虑，毕竟 Conv1 层的 size 还是比较大的，所包含的特征跟直接的图片像素信息也过于接近)。另外还对这些反映不同级别图片信息的各层自上向下进行了再处理以能更好地组合从而形成较好的特征表达（详细过程会在下面章节中进一步介绍)。而此方法正是我们本文中要讲的 FPN CNN 特征提取方法。

1、Feature pyramid Network

Input: 一张任意尺度的图片 Output: 多级特征预测的概率

Bottom-up pathway：是 backbone convnet 的前向计算。将每层的 layer 当成这里的 stage，将这些层的输出当成参考的 feature maps，会丰富我们定义的金字塔特征。这里使用 Resnets 作为 backbone，We denote the output of these last residual blocks as C2; C3; C4; C5 for conv2, conv3, conv4, and conv5 outputs, 这里没用 conv1，因为会占用较多的内存。

Top-down pathway and lateral connections(横向连接)：结构如下图所示

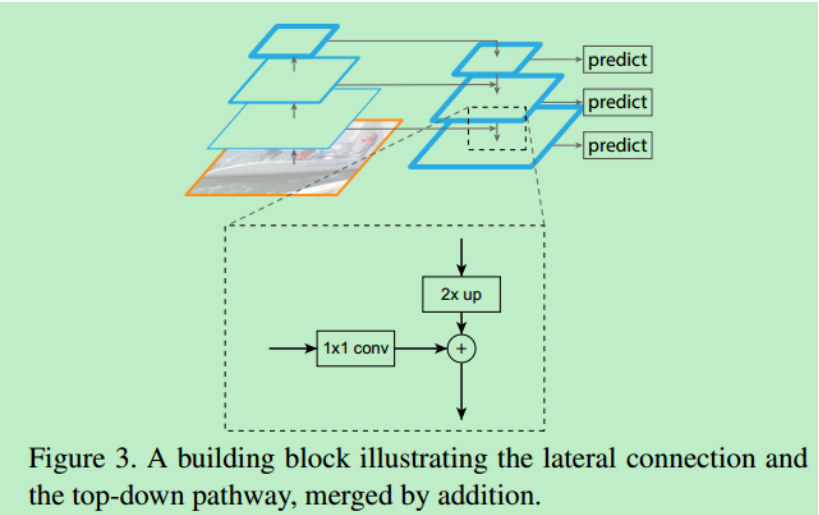

Figure 3. A building block illustrating the lateral connection and the top-down pathway, merged by addition.

图 2: Top-down pathway and lateral connections

top-down 和 bottom-up 一一对应，每一个横向连接都有相同大小的空间尺寸。bottom-up 包含低级的语义信息，但位置信息较精确。Finally, **we append a 3×3 convolution on each merged map** to generate the final feature map, which is to reduce the aliasing effect of upsampling.,This final set of feature maps is called P2; P3; P4; P5.

经过 classifiers/regressors 后，他们的输出都为 256-d，**Designing better connection modules is not the focus of this paper, so we opt for the simple design described above.**

2、Applications

we adopt our method in RPN for bounding box proposal generation and in Fast R-CNN for object detection.

(1) Feature Pyramid Networks for RPN

we assign anchors of ***a single scale*** to ***each level.***。这里总共有 15 个 abchors

anchors on a specific level. Instead, we assign anchors of a single scale to each level. Formally, we define the anchors to have areas of $\{32^2, 64^2, 128^2, 256^2, 512^2\}$ pixels on $\{P_2, P_3, P_4, P_5, P_6\}$ respectively.[1] As in [29] we also use anchors of multiple aspect ratios {1:2, 1:1, 2:1} at each level. So in total there are 15 anchors over the pyramid.

图 3: anchors

(2)Feature Pyramid Networks for Fast R-CNN

Fast RCNN 里，FPN 主要应用于选择提取哪一层的 feature map 来做 ROI pooling。假设特征金字塔结果对应到图像金字塔结果。定义不同 feature map 集合为 P2, P3, P4, P5，对于输入网络的原图上 w*h 的 ROI，选择的 feature map 为 Pk，其中 224 为 ImageNet 输入图像大小，下图为对应的计算公式

$$k = \lfloor k_0 + \log_2(\sqrt{wh}/224) \rfloor.$$

图 4:

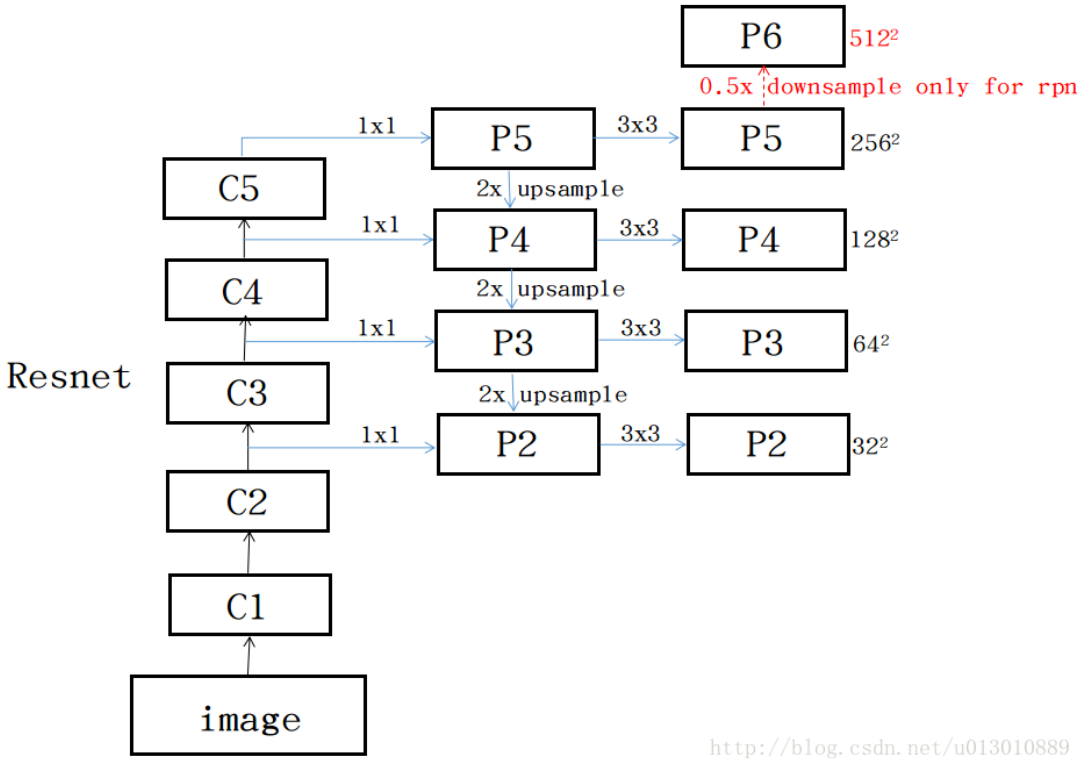

图 5: FPN 框架 1

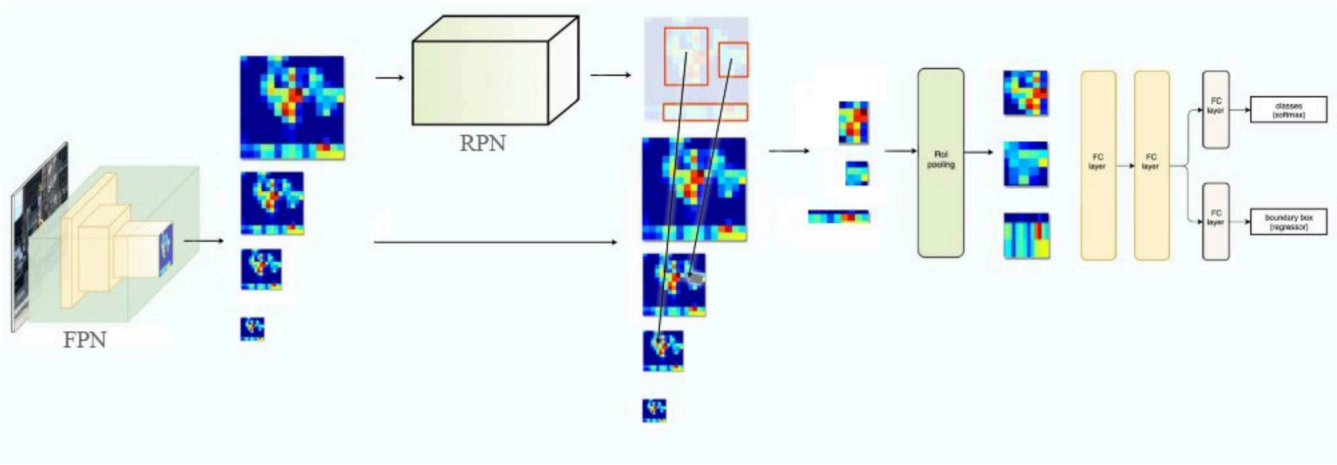
FPN
RPN
RoI pooling
FC layer
FC layer
FC layer
FC layer
classes (softmax)
boundary box (regressor)

图 6: FPN 框架 2